A rule based approach to classification of EEG datasets: a comparison between ANFIS and rough sets

Pari Jahankhani
Kenneth Revett
Vassilis Kodogiannis

Harrow School of Computer Science

A Rule Based Approach to Classification of EEG Datasets: A Comparison Between ANFIS and Rough Sets

Pari Jahankhani, Kenneth Revett and Vassilis Kodogiannis

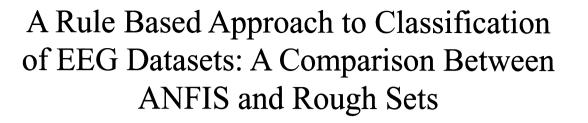*Abstract*—**This paper compares two different rule based classification methods in order to evaluate their relative efficiacy with respect to classification accuracy and the caliber of the resulting rules. Specifically, the application of Adaptive Neuro-Fuzzy Inference System (ANFIS) and rough sets were deployed on a complete dataset consisting of electroencephalogram (EEG) data. The results indicate that both were able to classify this dataset accurately and the number of rules were similar in both cases, provided the dataset was pre-processed using PCA in the case of ANFIS.**

Index Terms—**electroencephalography, Neuro-fuzzy systems, PCA, Rough sets, and wavelets**

I. INTRODUCTION

IN this study, a comparison of two rule-based classifiers was undertaken in order to investigate the classification accuracy and the resulting rule base. We compared the application of a neuro-fuzzy model (ANFIS) and rough sets, a by now classical rule-based classifier. Neuro-fuzzy systems have been deployed successfully in many applications, and yields a rule set that is derived from a fuzzy perspective inherent in data. The dataset employed in this study consisted of two short segments of continuous EEG recordings from normal and patients in the midst of an epileptic seizure. The EEG was collected over a short time interval from a 128 electrode montage, using standard mastoid references. The purpose of the classifiers was to provide an automated labeling method to identify epilepsy from EEG recordings.

The structure of this paper is as follows: we briefly describe the dataset, followed by any pre-processing stages (for ANFIS). Next the results of the application of rough sets and ANFIS are presented, followed by a brief discussion of the results of this study.

P. Jahankhani is with the School of Computer Science, University of Westminster, London HA1 3TP, UNITED KINGDOM (e-mail: parij@wmin.ac.uk).

K. Revett is with the School of Computer Science, University of Westminster, London HA1 3TP, UNITED KINGDOM (e-mail: revettk@wmin.ac.uk).

V. Kodogiannis is with the School of Computer Science, University of Westminster, London HA1 3TP, UNITED KINGDOM (e-mail: kodgiv@wmin.ac.uk).

II. DATA SELECTION AND RECORDING

We have used the publicly available data described in Andrzejak *et al.* [1]. The complete data set consists of two sets (denoted N for normal and E for eleptogenic), each containing, 128 channel EEG data, recorded over a short time interval (approximately 1 minute). These segments were selected and cut out from continuous multi-channel EEG recordings after visual inspection for artefacts, e.g., due to muscle activity or eye movements. Volunteers were relaxed in an awake-state with eyes open (N). Sets E originated from EEG archive of pre-surgical diagnosis. EEGs from five patients were selected, all of who had achieved complete seizure control after resection of one of the hippocampal formations, which was therefore correctly diagnosed to be the epileptogenic zone. Segments, set E only contained seizure activity.

Here segments were selected from all recording sites exhibiting ictal activity. All EEG signals were recorded with the same 128-channel amplifier system, using an average common reference. The data were digitised at 173.61 samples per second using 12 bit resolution. Band-pass filter settings were 0.53–40 Hz (12dB/oct). In this study, we used two dataset (N and E) of the complete dataset.

A. Pre-processing

In this work a Discrete Wavelet Transform has been used to extract features (see [2] for a discussion on wavelets). The number of decomposition levels was chosen to be 4. thus, the EEG signal were decomposed into the details D1-D4 and one final approximation, A4. The wavelet coefficients were computed using the db2 in this study. The following statistical features were used to represent the time- frequency distribution of the EEG signals.

- Maximum of the wavelet coefficients in each sub-band.
- Minimum of the wavelet coefficients in each sub-band.
- Mean of the wavelet coefficients in each sub-band
- Standard deviation of the wavelet coefficients in each sub-band

Principal component analysis (PCA) was used to reduce the data dimensionality for the ANFIS system only (see [3] for a discsion on PCA).

III. ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM ARCHITECTURE

In this section, we describe ANFIS. ANFIS has proven to be excellent approximation tool. ANFIS implements a first order Sugeno model in a framework of a five-layer adaptive network [4]. The ANFIS architecture is a common rule set with two fuzzy if-then rules is the following:

Rule 1: If x is A_1 and y is B_1, then $f_1=p_1 x+ q_1 y +r_1$

Rule 2: If x is A_2 and y is B_2, then $f_2=p_2 x+ q_2 y +r_2$
are the fuzzy sets, f_i are the outputs within the fuzzy region

Layer 1: every node I in this layer is an adaptive node with a node function

$$O_{1,i} = \mu A_i(x), \quad \text{for } i=1,2 \text{ or}$$

$$O_{1,i} = \mu B_{i-2}(y), \quad \text{for } i=3,4$$

Where x (or y) is the input to node i and A_i (or B_{i-2}) is a linguistic label associated with this node, $\mu A_i(x)$, $\mu B_{i-2}(y)$ can adopt any fuzzy membership function, such as the generalized bell shape function :

$$\mu A(x) = \frac{1}{1+ \left| \dfrac{x - c_i}{a_i} \right|^{2b_i}}$$

where a_i, b_i and c_i are the parameters of the membership function. As the value of these parameters change the bell-shaped function varies accordingly. Parameters in this layer are referred to as premise parameters.

Layer 2: Every node in this layer is fixed node labelled p, whose output is the product of all the incoming signals. The outputs of this layer can be represented as :

$$O_{2,i} = w_i = \mu A_i(x)\mu B_i(y), \text{ i=1,2}$$

Each node output represents the firing strength of a rule. In general, any other T-norm operators that perform fuzzy AND can be used as the node function in this layer.

Layer 3: Every node in this layer is a fixed node labelled N. The i-th node calculates the ratio of the i-th rules firing strength to the sum of all rules firing strengths.

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, \text{ i=1,2}$$

Output of this layer is called normalized firing strengths.

Layer 4: Every node i in this layer is an adaptive node with a node function

$$O_{4,i} = \overline{w}f_i = \overline{w}(p_i x+q_i y+r_i)$$

Where $\overline{w}$ is a normalized firing strength from layer 3,

and $\{p_i, q_i, r_i\}$ is the parameter set. These parameters are called consequent parameters [11],[12].

Layer 5: The single node in this layer is labeled Σ that computes the overall output as the summation of all incoming signals.

$$O_{5,1} = \sum_i \overline{w_i}f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}$$

The adjustment of modifiable parameters is a two step process, first, information is propagated forward in the network until layer 4, where the parameters are identified by a least-squares estimator. Then the parameters in layer 2 are modified using gradient descent.

IV. ROUGH SETS

Rough set theory is a relatively new data-mining technique used in the discovery of patterns within data first formally introduced by Pawlak in 1982 [5]. Since its inception, the rough sets approach has been successfully applied to deal with vague or imprecise concepts, extract knowledge from data, and to reason about knowledge derived from the data. We demonstrate that rough sets has the capacity to evaluate the importance (information content) of attributes, discovers patterns within data, eliminates redundant attributes, and yields the minimum subset of attributes for the purpose of knowledge extraction.

The first step in the process of mining any dataset using rough sets is to transform the data into a decision table. In a decision table (DT), each row consists of an observation (also called an object) and each column is an attribute, one of which is the decision attribute for the observation. In our case the decision table consists of 8000 rows and each row contains a vector of 20 numbers (values of 20 numerical conditional attributes) labelled with one of 5 decision values (N & E). Objects that share the same decision value are said to belong to one *decision class*. Attributes other than decision will be referred to as *conditional attributes* or simply *conditions*.

Luckily enough, thanks to the fact that data table is generated by a controlled pre-processing algorithm, we have no missing values and no errors which is not commonplace in medical data sets. The table is consistent, i.e. there are no two rows that have the same conditional part and different decisions.

One of most characteristic features of our data set is that all conditional attributes are numeric (floating point numbers). In order to apply some of the RS methods to such data table one have to perform *discretisation*.

Discretisation refers to partitioning attributes into intervals – tantamount to searching for "cuts" in the range of attribute. All values that lie within a given range (between two cuts) are mapped onto the same value, transforming interval into categorical data. In this study we apply discretisation method based on Maximal Discernibility

(MD) heuristics that makes use of the core RS notion of *discernibility* between decision classes. Details of this method and its implementation in RSES by way of example are given in [6].

The ultimate goal we want to achieve with our RS toolkit is to construct a *classifier* – that is a procedure which when given an unlabelled object is capable of assigning a proper decision value. In particular, we will be dealing with classifiers that are based on *decision rules*.

Decision rule *r* is a formula:

$$(a_i=v_i)\square \ ... \ \square(a_i=v_i)\square d=v_d$$

where atomic sub-formula $a_i=v_i$ is called *descriptor* or *condition*. We say that rule *r* is *applicable* to an object, or alternatively, the object *matches* rule, if its attribute values satisfy the premise of this rule. With the rule we can associate some numerical characteristics derived from the underlying data table DT . *Support(r)* is equal to the number of objects from table for which rule *r* applies correctly, i.e., the premise of rule is satisfied and the decision given by rule is similar to the one preset in decision table. *Matching(r)* is the number of objects in the table for which rule *r* applies in general. Analogously the notion of matching set for a rule or collection of rules may be introduced (see [6]).

V. RESULTS

The dataset was used without any pre-processing with rough sets – the decision classes was set to either N or E (equal numbers of both classes was used for training – in a 70/30 split). The results from the application of rough sets to this dataset are presented in table 1.

TABLE I
THE NUMBER OF RULES AND CLASSIFICATION ACCURACY AS A FUNCTION OF THE ATTRIBUTE SETS WHEN ROUGH SETS WAS EMPLOYED AS THE CLASSIFICATION ALGORITHM

Attribute	No. of rules	Accuracy	Attribute	No. of rules	Accuracy
Max. D2	2	99.8%	Min. D4	2	99.9%
Min. D2	10	99.6%	Max D4	2	100%
St. dev. D2	14	99.2%	Min. A4	6	99.9%
Min D3	8	99.9%	Max A4	10	99.7%
Max D3	2	99.9%	St. dev. A4	2	99.9%
St. dev. D3	2	99.8%			

Note that the result of applying rough sets yielded a rule set which contained a single antecedent as describe in the following:

If MaxD4 < 173.1838 then Decision = N;
If MaxD4 > 173.1838 then Decision = E

where MaxD4 is an attribute derived from the data utilizing the DWT. Note that the threshold value of 173.1838 was derived from the rough sets generate rule.

The application of ANFIS as the rule-base classifier is presented in tables 2 & 3. Note that prior to the application of ANFIS, the dataset was reduced in terms of the number of attributes both by the application of PCA (and selecting the principal components that comprised 95+ % of the data). To compare the classification results of both ANFIS and rough sets, the attribute selected by rough sets (MaxD4) and the first principal component were used as the input into the ANFIS system.

TABLE II
THE CLASSIFICATION ACCURACY USING ANFIS WITH EITHER THE ROUGH SETS SELECTED ATTRIBUTE (MAXD4 OR THE 1ST PRINCIPAL COMPONENT FOUND BY THE APPLICATION OF PCA). EITHER A 2 OR 3-PARTITION WAS USED IN THE APPLICATION OF THE FUZZY RULE SET (ANFIS)

Input	Rule No.	Class N	Class E
MaxD4	2	100%	97.3%
MaxD4	3	100%	97.7%
PCA 1 Component	2	100%	98.8%
PCA 1 Component	3	98%	100%

TABLE III
THIS TABLE PRESENTS THE RESULTS OF THE APPLICATION OF ANFIS USING EITHER: A) THE MAXD4 ATTRIBUTE, B) THE 1ST PRINCIPAL COMPONENT, AND C) THE 1ST AND 2ND COMPONENT IN TERMS OF CLASSIFICATION ACCURACY USING A 70/30 TRAINING ALGORITHM

Input	No. Rules	Class N	Class E
MaxD4	2	100%	97.3%
MaxD4	3	96.4%	97.7%
PCA 1 Component	2	98.2%	98.8%
PCA 1 Component	3	98%	100%
PCA 2 Components	4	99%	98.8%
PCA 2 Components	9	89.8%	98.6%

VI. CONCLUSIONS

The results from this study indicate that both the ANFIS and rough sets approach were able to classify an EEG dataset containing samples from normal and eleptogenic patients with virtually 100% accuracy. The principle difference between these two rule-based approaches is in the data pre-processing stages. ANFIS works well with small datasets in terms of the number of features (typically 6 or so is the maximum), rough sets typically can work with datasets with a much larger number of features, provided there are sufficient numbers of examples for each decision class (a typical heuristic is 100:1). With ANFIS, when the full set of features (20) was deployed, the system typically crashed before the rule set was generated. Therefore, in order to deploy ANFIS to datasets with a large number of attributes, some sort of dimensionality must be performed. PCA was a suitable choice in this study, resulting in a classification accuracy (see [6] for a discussion of how classification accuracy is measured) approaching that of rough sets.

Whether or not PCA is an appropriate dimensionality reduction technique is dependent on the dataset. In this study, deploying the 1st principle component (accounting for approximately 95% of the variance) yielded a higher classification accuracy than deploying the 1st two principal components (accounting for 99% of the variance). This result might indicate that variance is not the best selection criterion for this type of dataset, and may indicate that other dimensionality techniques might need to be considered. This remains an open question that may require empirical analysis in order to determine what method is best for a particular class of data. This is a topic for future research.

REFERENCES

[1] Andrzejak R. G., Lehnertz K., Mormann F., Rieke C., David P., Elger C. E.,Indications of nonlinear deterministic and finite-dimensional structures in timeseries of brain electrical activity. *Phys. Rev. E*, 64, pp. 1-8, 2001.

[2] Daubechies, I., orthonormal bases of compactly supported wavelets, Comm. Pure App.Math.41(7)(1988) 909-9996

[3] Jolliffe, I.T. Principal Component Analysis. Springer, New York, NY, 1986

[4] Jang J.S.: ANFIS: Adaptive-Network-based Fuzzy Inference System. IEEE Trans. On System, Man and Cybernetics. Vol.23, No.3, May/June 1993, PP.665-685

[5] Pawlak, Z., Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

[6] Revett, K., Szczuka, M., Jahankhani, P., & Kodogiannis, V. Attribute Selection for EEG Signal Classification Using Rough Sets and Neural Networks, RSCTC2006, November, 2006, Kobe Japan , pp 408-417